Original Investigations

Research Paper ■

Retrieval Feedback in MEDLINE

PADMINI SRINIVASAN, PHD

Abstract **Objective:** To investigate a new approach for query expansion based on retrieval feedback. The first objective in this study was to examine alternative query-expansion methods within the same retrieval-feedback framework. The three alternatives proposed are: expansion on the MeSH query field alone, expansion on the free-text field alone, and expansion on both the MeSH and the free-text fields. The second objective was to gain further understanding of retrieval feedback by examining possible dependencies on relevant documents during the feedback cycle.

Design: Comparative study of retrieval effectiveness using the original unexpanded and the alternative expanded user queries on a MEDLINE test collection of 75 queries and 2,334 MEDLINE citations.

Measurements: Retrieval effectivenesses of the original unexpanded and the alternative expanded queries were compared using 11-point-average precision scores (11-AvgP). These are averages of precision scores obtained at 11 standard recall points.

Results: All three expansion strategies significantly improved the original queries in terms of retrieval effectiveness. Expansion on MeSH alone was equivalent to expansion on both MeSH and the free-text fields. Expansion on the free-text field alone improved the queries significantly less than did the other two strategies. The second part of the study indicated that retrieval-feedback–based expansion yields significant performance improvements independent of the availability of relevant documents for feedback information.

Conclusions: Retrieval feedback offers a robust procedure for query expansion that is most effective for MEDLINE when applied to the MeSH field.

■ **JAMIA.** 1996;3:157–167.

The aim of this study was to evaluate methods for designing effective queries for searching the MEDLINE database. It was motivated by the need to improve upon MEDLINE search performance as established in previous investigations by the medical informatics community.

MEDLINE users have to be aware of two different vocabularies when specifying queries. The first is the controlled vocabulary underlying the MeSH indexes of documents. The second is the free-text vocabulary underlying search fields such as document titles and abstracts. These two vocabularies differ significantly. MeSH concepts are selected by trained indexers from a controlled vocabulary, while titles and abstracts are

Affiliation of the author: Department of Computer Science, Cornell University, Ithaca, NY.

Correspondence and reprints: Padmini Srinivasan, PhD, School of Library and Information Science, University of Iowa, 3067 Main Library, Iowa City, IA 52242.
e-mail: padmini-srinivasan@uiowa.edu

Received for publication: 7/25/95; accepted for publication 11/13/95.

generally written by authors in free-text format. Other differences include size, linguistic structure, and specificity of component terms. End-users are typically unskilled in the use of MeSH and are therefore often unable to specify proper MeSH search criteria.[1] Thus, the MeSH field, an important indicator of text content,[2] tends to be underexploited during searching. Free-text search fields also pose certain challenges to users. For example, they contain larger numbers of general search terms. General terms when used in queries tend to decrease search accuracy. Query-expansion research focuses on the design of strategies that assist the user in query formulation. In particular, the objective in query expansion is to automatically modify a user's original query into one that is more effective in retrieval. For example, a MEDLINE query-expansion strategy may operate by adding query terms selected from MEDLINE's free-text and MeSH vocabularies.

This study investigated *retrieval feedback* as a mechanism for the automatic expansion of queries. Retrieval-feedback–based query expansion refers to the process by which a user's original query is utilized to conduct an initial retrieval run on the document collection. Information from the top few documents retrieved by this initial run is used to modify the original query.

The first objective of this study was to compare three different expansion strategies developed within the retrieval-feedback framework. The first expansion strategy aimed to improve upon a query's MeSH-search criteria. (Initial exploration of this strategy was conducted in an earlier pilot study.[3]) The second strategy was designed to improve upon a query's free-text search criteria. Finally, the third query-expansion strategy sought to improve upon both sets of search criteria. The second objective of this study was to examine the dependence of retrieval-feedback–based query expansion on the availability of relevant documents for feedback information.

Background

Query Expansion for MEDLINE

Query expansion is a vibrant research focus in information retrieval. A number of general query-improvement strategies have been designed, both for MEDLINE and for databases in other subject domains. Terms (words and phrases) may be added to a user's initial query using either semantic[1,4–6] or statistical approaches.[7–14] One alternative is to add terms from relevant documents through a process called relevance feedback.[15–18] Retrieval feedback is a derivative of relevance feedback where the documents used for feedback information need not be actually relevant to the user.[3,19] Recent query-expansion research paying specific attention to MEDLINE is briefly described here.

Hersh et al. tested a novel approach of using the titles and abstracts of documents to extract concepts from the metathesaurus component of the UMLS (Unified Medical Language System) product.[1] This metathesaurus is a comprehensive vocabulary that combines information from a variety of source vocabularies, a dominant one being MeSH.[20] The authors used an experimental retrieval system called SAPHIRE to index both the documents and the queries with metathesaurus concepts. Following this, SAPHIRE's phrase-matching algorithm was used to retrieve documents. The authors discovered that SAPHIRE performed similarly to both novice physicians and expert physicians. However, they found significant differences between SAPHIRE's performance and the performance of librarians.

Aronson et al. explored a method based on "underspecified syntactic analysis" for identifying appropriate concepts for indexing from the UMLS metathesaurus.[4,5] Their technique is similar to the robust method for mapping text to MeSH concepts previously proposed and investigated by Elkin et al.[21] To the best of this author's knowledge, Elkin et al. have not tested the impact of their mapping method on retrieval. Both groups identify concepts from the simple noun phrases in texts. After identifying these noun phrases, Aronson et al. employ a comprehensive program for intensive variant generation.[4,5] For example, the program considers abbreviation expansion and derivational morphology. Metathesaurus concepts containing any variants are evaluated, and the qualifying concepts substitute the noun phrases in the original texts. Aronson et al. tested this approach with SMART[22] as their retrieval system using a test collection of 3,000 MEDLINE documents and 150 queries in three subject areas. The authors report an increase of 4% in average precision when compared with retrieval using unmodified queries and documents.[4]

Yang and Chute investigated MeSH query construction as part of a larger investigation to solve for vocabulary differences between the queries and the documents.[7–9] They used mappings from query words to MeSH concepts for query expansion. In their 1993 study[9] these were derived using the linear least-square method, while in 1994[7] they used an expert network to generate the mappings. In either case, their method requires a training set of example queries and their relevant documents. Test queries expanded using these learned mappings improve retrieval performance. When the mappings were applied to a subset of the MEDLINE collection used by Hersh et al.,[1] as well as in this study, Yang and Chute observed a 32.2% improvement in average precision over their baseline of 0.412.[7,9] The Yang and Chute studies[7-9] certainly emphasize the observation that MeSH terms are important for retrieval. A practical disadvantage of their approach is its reliance on the availability of a training set of queries and relevant documents, which limits the applicability of their method. More important, since 88% of their test queries appear in their training set, they make the questionable assumption that a new query is likely to be similar to at least some of the queries in the training set. Since performance data for the nonrepeating 12% of test queries are not provided, it is not possible to predict what might happen in realistic situations, where this assumption will most likely be violated.

Query Expansion Using Retrieval Feedback

In relevance feedback, a query is modified using information in previously retrieved documents that have been judged for relevance by the user.[15-18] Methods such as Rocchio's[17] and Ide's,[16] which differ in their query-expansion details, have been studied within this broad strategy. By design, relevance feedback relies on the availability of relevance decisions by users. Such decisions are unlikely to be available in most ad hoc querying environments. Hence, an alternative that has recently been explored is to expand the query using a relevance-independent feedback process, which we refer to as retrieval feedback.[19] Here, an initial retrieval run is conducted and the top few documents retrieved are all assumed to be relevant and are used for query expansion.

Retrieval feedback works very well in the full-text TREC experimental environment, producing improvements of as much as 20% over baseline precision scores as shown by the Cornell group using SMART.[19] The MEDLINE database is very different from the TREC collection. In the MEDLINE test collection, the queries are typically smaller; documents are not full-text; the number of relevant documents per query is smaller, 14 for the test collection used here as opposed to 282 for TREC. Thus, our first goal was to determine whether retrieval feedback was also effective in automatically expanding MEDLINE queries. Our second goal was to investigate factors that determine the success of the expansion method. These goals are specified next.

Current Research Questions

We posed the following two questions in this study:

Question 1: Can retrieval feedback be used to effectively expand the original query? Three expansion strategies were investigated. The first strategy sought to improve upon the MeSH-text search criteria of the query. The second strategy improved upon the free-text search criteria. Finally, the third alternative strategy aimed to improve upon the query's free-text and MeSH search criteria.

Question 2: Does query expansion based on retrieval feedback have any implicit dependencies on the inclusion of relevant documents in its feedback set? This research question sought to examine any implicit reliance of the feedback technique on the presence of items that were actually relevant in the document set used for feedback information.

Methods

Test Environment

The experiments were run using Cornell's SMART retrieval system,[22,23] a sophisticated and powerful research system based on the vector-space model and designed for testing ideas on information retrieval.

MEDLINE Test Collection

The test collection of 75 queries and 2,344 MEDLINE documents produced by Hersh et al.[1] was used for this study. All queries had some relevant documents in the collection. All documents included abstracts.

Indexing Strategies

In SMART, documents and queries are automatically indexed to yield a weighted vector of index terms (words or phrases), as shown below for a document D_1. Term weights reflect the relative importances of the terms when representing the document.

$$D_1 = (wt_{D_1}t_1, wt_{D_1}t_2, \ldots, wt_{D_1}t_m)$$

where $wt_{D_1}t_i$ represents the weight of term t_i for document D_1 and an indexing vocabulary of size m is assumed. SMART conducts retrieval by computing similarity as the vector inner product of the document and the query vectors, resulting in a ranked list of documents for a given query.

Two word-based index vectors were derived for each document, a vector from the nontrivial words in the title and abstract (ta-vector) and a vector from the

Table 1 ■

Term-weighting Strategies in SMART

Dimension	Symbol	Interpretation
A	b: binary	1 if term is present and 0 if term is absent in document
A	a: augmented	(0.5 + 0.5 × *tf/max_tf_in_document*)
A	l: logarithmic	1 + ln(tf)
A	n: none	tf
B	t	ln(N/n); N = no. docs. in database; n = no. docs. with term.
B	n	idf not used
C	c: cosine	$\sqrt{w_{D_1}c_1^2 + \cdots + w_{D_1}c_m^2}$; m = vocabulary size
C	n	no normalization

nontrivial words of the MeSH concepts (m-vector). (For ease of distinction we henceforth refer to words of the titles and abstracts as *words* and words of the MeSH concepts as *concepts*.) Assuming a title-abstract vocabulary of p words and a MeSH vocabulary of q concepts, a document may be represented as:

$$D_1 = (wt_{D_1}w_1, wt_{D_1}w_2, \ldots, wt_{D_1}w_p); \quad (wt_{D_1}c_1, wt_{D_1}c_2, \ldots, wt_{D_1}c_q) \qquad (1)$$

We generated a single ta-vector for each query, as we considered the user's initial free-text query more suitable for searching the title and abstract field.

$$Q_1 = (wt_{Q_1}w_1, wt_{Q_1}w_2, \ldots, wt_{Q_1}w_p) \qquad (2)$$

SMART allows a wide variety of strategies for computing term weights. Each strategy is represented by a triple: ABC. A represents the term frequency component, *i.e.,* the number of times the term occurs in the document. B represents the inverse document frequency component, which increases with the rarity of the term in the database. C represents the normalization component for the length of the document. Options are available for each of these three dimensions of the triple, as shown in Table 1.

Retrieval Strategies

A retrieval strategy in SMART is defined by the weighting strategies used to represent documents and queries. Thus, it is represented by a pair of triples, e.g., atc.atn, where the first triple and the second triple represent the document and the query-weighting strategy, respectively. Given the indexing options in Table 1, there are 256 (16 × 16) retrieval strategies that may be explored within SMART. We used atn.atc and atc.atc, two highly effective strategies identified from our pilot study,[3] as our baseline retrieval strategies.

Since SMART represents both documents and queries by weighted vectors, retrieval is conducted by computing the similarity between every query and document pair. Thus, every query–document pair yields a numerical similarity value representing the closeness between the two entities. SMART uses these similarity values to rank the entire database for a given query. (This is in contrast to standard Boolean retrieval systems, which for a given query partition the database into two sets: documents that are retrieved and documents that are not retrieved.) Given that SMART ranks all documents of the database by query similarity, the retrieved result, i.e., items shown to the user, consists of all documents above a threshold rank or a threshold similarity value. This threshold may be set by the user.

Evaluation of Retrieval Strategies

Since SMART retrieves documents by ranking them against queries, alternative retrieval strategies were compared based on ranking effectiveness, i.e., ability to rank relevant documents in the database higher than nonrelevant ones. The 11-AvgP measure used here stands for 11-point-average precision and was designed to evaluate ranked sets of documents. Recall is the proportion of relevant documents retrieved, while precision is the proportion of the retrieved documents that is relevant. Given a ranked set of documents, precision may be computed at the 11 standard recall points of 0%, 10%, . . . , 100%. The final precision score of a retrieval strategy at a standard recall point is the average of precision scores at that point computed for each test query. This averaging technique yields *macro average* data wherein each test query is allowed to contribute equally to the overall performance score for the system.[23,page 538]

Alternative Query Expansion Strategies

Given a user's initial query Q_1 represented by a single ta-vector (equation 2): our objective was to expand Q_1 into a new query Q_2 using three alternative expansion strategies, which are described next.

- Expansion strategy 1 (ES:ta'-m'):

  This is the most general expansion strategy, where the original ta-vector is expanded and a new m-vector of MeSH concepts is added to form Q_2.

$$Q_2 = (wt_{Q_2}w_1, wt_{Q_2}w_2, \ldots, wt_{Q_2}w_p); \quad (wt_{Q_2}c_1, wt_{Q_2}c_2, \ldots, wt_{Q_2}c_q) \qquad (3)$$

- Expansion strategy 2 (ES:ta-m'):

  Here the original ta-vector remains the same. Only a new m-vector is added to form Q_2.

$$Q_2 = (wt_{Q_1}w_1, wt_{Q_1}w_2, \ldots, wt_{Q_1}w_p); \quad (wt_{Q_2}c_1, wt_{Q_2}c_2, \ldots, wt_{Q_2}c_q) \qquad (4)$$

- Expansion strategy 3 (ES:ta'):

  Here Q_2 has no m-vector. Instead, only the ta-vector is expanded to give a new Q_2.

$$Q_2 = (wt_{Q_2}w_1, wt_{Q_2}w_2, \ldots, wt_{Q_2}w_p) \qquad (5)$$

Identifiers ES:ta'-m', ES:ta-m' and ES:ta' are used to represent expansion strategies 1, 2, and 3, respectively. A ta' (or m') symbol indicates that the final ta-vector (or m-vector) is obtained through expansion. Obviously, there is no expansion strategy involving a plain m, since there is no original m-vector in the query. Retrieval on expanded queries is conducted by computing similarity as a weighted sum of the inner products of corresponding vectors in documents and queries.

$$Similarity(D, Q) = delta \times similarity(ta\text{-}vectors) + similarity(m\text{-}vectors) \qquad (6)$$

where delta is a parameter that allows one to change the relative emphases on the two types of vectors during retrieval.

The SMART system is designed to support relevance-feedback runs. The only slight modification that is required for retrieval feedback is to designate each document in the feedback set as relevant. Ide's method of feedback was used in this study.[16] In this method, the weight for a term t_i in the old query is modified as:

$$wt_{Q_{new}}t_i = alpha \times wt_{Q_{old}}t_i + beta \times \sum_{k=1}^{R}(wt_{Doc_k}t_i) - gamma \times wt_{DocNR_1}t_i \qquad (7)$$

where $wt_{Q_x}t_i$ is the weight of term t_i in query Q_x, R is the number of relevant documents retrieved, and NR_1 is the topmost nonrelevant document in the feedback set. Alpha represents the relative importance of term t_i in the original query. Beta and gamma are parameters designating the relative importance of relevant information vs nonrelevant information for query expansion. Since all documents used for retrieval feedback are assumed to be relevant, the negated term in equation 7 drops off. Moreover, when adding concepts to m-vectors, the first term also drops off, since there are no original m-vectors in the queries. When adding new free-text terms to a query, alpha and beta were always kept equal to each other in these experiments.

Ide's method of feedback involves two parameters. The first parameter controls the size of the feedback set of documents. We set this at 3, 5, 10, 15, and 20 documents. The second parameter is the number of new words or concepts to be added to a vector, and was set at 5, 10, 15, and 20. Also, the values for these parameters were kept the same across both vectors for the ES:ta'-m' strategy.

There are two steps to accomplish for query expansion. First, appropriate words and concepts have to be selected for the query. Second, their weights have to be computed. All words and concepts in the feedback set of documents for a query were considered to be candidates for addition to the corresponding vectors, depending upon the expansion strategy. These were ranked by the measure in equation 7 and items were selected from the top of this list according to the parameters governing the numbers of words and concepts to add to the vectors. The final weight assigned to an added word or concept is also specified by the same equation.

An Example

The following example illustrates the query expansion process for a single query.

Text of User Query: "Patient with mycosis fungoides, wishes to assess treatment options."

Values of different options used in this example:

Retrieval strategy: atn.atc

Expansion strategy: ES:ta'-m'

Size of feedback set: 15

Number of terms added to each query vector: 10

Delta value (see equation 6): 1.25

First, the above textual query is preprocessed using SMART to generate the query ta-vector displayed in Table 2. Query weights are computed using the atc indexing strategy. Notice that the original query words have been stemmed as part of the preprocessing by SMART. As mentioned before, the original query does not generate an m-vector.

Next, the query shown in Table 2 is used to conduct an initial retrieval run. That is, the documents of the database are ranked by similarity to the query's initial ta-vector. Table 3 shows the titles of the 15 top-ranked documents that form the feedback set used for query expansion.

Table 2 ■

Original Query Representation: ta-vector with atc Indexing

Term	Weight
option	0.70459
fungoid	0.45910
mycos	0.45428
assess	0.27158
treat	0.09979
patient	0.05189

Table 3 ■

Top 15 Documents Retrieved by Original Query

Document No.	Document Title
234	Photochemotherapy (PUVA) in the pretumor stage of a mycosis fungoides: report from the Scandanavian Mycosis Fungoides Study Group
8	Bleomycin therapy in mycosis fungoides
1390	Cutaneous malignancies and metastatic squamous cell carcinoma following topical therapies for mycosis fungoides
122	Preliminary evaluation of 15 chemotherapeutic agents applied topically in the treatment of mycosis fungoides
244	Demethylchlortetracycline and griseofulvin as examples of specific treatment for mycosis fungoides
108	Prednimustine in mycosis fungoides: a report from the Scandinavian Mycosis Fungoides Study Group
106	Mycosis fungoides plaque stage treated with topical nitrogen mustard with and without attempts at tolerance induction: report from the Scandinavian Mycosis Fungoides Study Group
1942	Successfully treated Hodgkin's disease followed by mycosis fungoides: case report and review of the literature
72	Total skin electron irradiation in mycosis fungoides
1179	Surgical stabilization of pathological neoplastic fractures
1725	Postmenopausal osteoporosis
310	Treatment of mycosis fungoides: total-skin electron-beam irradiation vs topical mechlorethamine therapy
104	Treatment of mycosis fungoides with heat-killed BCG and cord factor
31	Combined chemotherapy (COP) in treatment of mycosis fungoides: report of four cases
1719	Combined total body electron beam irradiation and chemotherapy for mycosis fungoides

Table 4 ■

Final Query Representation: Expanded ta-vector, New m-vector

ta-vector Word Weight	ta-vector Word	m-vector Concept Weight	m-vector Concept
0.47778	fungoid	0.32725	fungoid
0.47276	mycos	0.31810	mycos
0.21939	topic	0.27783	skin
0.18527	remit	0.16011	neoplasm
0.18400	complet	0.08561	therap
0.17598	stag	0.07527	age
0.10839	therap	0.07001	middl
0.10717	tumor	0.05559	male
0.09849	treat	0.04432	femal
0.09673	option	0.01086	human
0.08683	report		
0.07248	month		
0.06882	case		
0.06450	diseas		
0.05272	patient		
0.00276	assess		

Term information in the feedback set is used to modify the original query into its final form, shown in Table 4. Ten new terms have been added to the ta-vector, while a new m-vector of ten terms has been created. It may be observed that the term 'option,' with the highest weight in the original query, now ranks tenth in the expanded query's ta-vector, with the weight of 0.09673. Also, the term 'topic' (from the word 'topical' as in 'topical application'), which does not occur in the original query, now ranks third in the final query's ta-vector. Finally, the overlap between entries in the final ta- and m-vectors is low; only three terms are in common: therap, mycos, and fungoid. This emphasizes the point that although expansions on ta- and m-vectors utilize the same feedback set of documents, they occur independently.

Table 5 shows the change in ranks for this query's 32 known relevant documents when moving from the original query to the new expanded query created using the ES:ta'-m' expansion strategy. It may be seen that for 21 relevant documents the ranks improve using the expanded query. For ten relevant documents the ranks worsen, while one relevant document undergoes no change in rank. Interestingly, the ranks for two relevant documents (124 and 1108) improve dramatically. Due to this dramatic change, a cut-off rank of 54 is sufficient to retrieve all 32 relevant documents when using the expanded query. In contrast, a cut-off rank of 594 is needed when using the original query!

Appendix A provides an example of one document found relevant and Appendix B provides an example of a document found nonrelevant to this example query. Both free-text and MeSH fields of these documents are displayed.

Results

Table 6 presents the performance data for the three alternative query-expansion strategies. The table also provides details regarding the expansion strategies. For example, the first row states that the expansion strategy ES:ta'-m' yields a final 11-AvgP of 0.5907. The top 15 documents retrieved in the initial retrieval run form the feedback set, ten words are added to the ta-vector, and a new m-vector containing ten concepts is generated. The performance of this expanded query is 14.3% better than the performance of the original, unexpanded query. The atn.atc retrieval strategy is used for both the original and the expanded queries of this row. Finally, the value of delta is 1.25 when computing document-query similarity with the expanded query (see equation 6). (Notice that the example detailed in the previous section is from this first row of Table 6).

The six expansion-retrieval combinations of Table 6 provide significant improvements over corresponding baseline strategies ($p < 0.01$). The non parametric Wilcoxon signed-rank test for matched samples is used here. Notice that all expansion strategies have the same starting point *i.e.,* the original free-text query provided by the user. Therefore the differences in performance improvements may be attributed to differences in the final query vectors that are generated.

The expansion strategies are ranked, ES:ta-m', ES:ta'-m' followed by ES:ta'. ES:ta-m' gives a 1.3% improvement over ES:ta'-m' which in turn gives a 5.75% improvement over ES:ta-m. The first observation that can be made is that both ES:ta-m' and ES:ta'-m' strategies are significantly superior to the ES:ta' expansion strategy ($p < 0.02$). ES:ta', which ignores the MeSH field completely, does yield significant improvements over the baseline; however, these improvements are significantly less than those obtained by the other two strategies, which consider MeSH.

Table 5 ■

Comparing Ranks of Relevant Documents with Original and Final Queries

Relevant Document No.	Rank Using Original Query (Similarity Value)	Rank Using Final Query (Similarity Value)	Difference in Ranks: Original Rank − Final Rank
8	2 (3.6272)	5 (9.1220)	−3
31	14 (3.0236)	20 (7.3609)	−6
72	9 (3.1827)	17 (7.7017)	−8
83	42 (2.4290)	14 (7.8707)	28
96	22 (2.7362)	23 (6.6695)	−1
102	25 (2.7305)	36 (6.1156)	−11
106	7 (3.2765)	3 (9.2946)	4
109	41 (2.4309)	27 (6.4313)	13
123	36 (2.5048)	24 (6.5961)	12
124	594 (0.0752)	53 (2.6840)	541
125	33 (2.5783)	9 (8.2531)	24
134	21 (2.7544)	18 (7.6178)	3
154	44 (2.3529)	25 (6.5692)	19
234	1 (3.6368)	2 (9.4590)	−1
237	24 (2.7305)	21 (7.1462)	3
281	26 (2.7164)	34 (6.1489)	−8
310	12 (3.0479)	4 (9.2892)	8
471	47 (2.2575)	33 (6.1693)	14
479	39 (2.4379)	26 (6.5603)	13
606	23 (2.7309)	32 (6.1967)	−9
688	50 (2.0754)	41 (5.3152)	9
748	48 (2.2084)	47 (4.9004)	1
754	40 (2.4360)	40 (5.5807)	0
1044	34 (2.5721)	37 (5.7212)	−3
1108	541 (0.0776)	49 (3.8925)	492
1390	3 (3.6215)	1 (9.5959)	2
1399	53 (2.0077)	29 (6.3233)	24
1592	31 (2.6009)	28 (6.3803)	3
1692	54 (1.9694)	42 (5.2767)	12
1719	15 (2.9086)	10 (8.0519)	5
1760	18 (2.8496)	22 (6.6926)	−4
1763	46 (2.3318)	31 (6.2849)	15

The second observation that holds is that ES:ta-m' and

Table 6 ■

Performances of Alternative Query-expansion Strategies

Feedback Strategy	Retrieval Strategy	Baseline 11-AvgP	Expanded Query 11-AvgP	Size of Feedback Set	Vector Size	Delta
ES:ta'-m'	atn.atc	0.5169	0.5907 (+14.3%)	15	10	1.25
ES:ta'-m'	atc.atc	0.5027	0.5941 (+18.2%)	5	20	1.0
ES:ta'-m'	atn.atc	0.5169	0.6018 (+16.4%)	10	20	0.66
ES:ta'-m'	atc.atc	0.5027	0.5935 (+18.1%)	5	20	0.80
ES:ta'	atn.atc	0.5169	0.5619 (+8.7%)	15	20	—
ES:ta'	atc.atc	0.5027	0.5602 (+11.4%)	10	20	—

ES:ta'-m' yield equivalent final 11-AvgP scores. That is, whether combined with the query's original ta-vector or combined with an expanded ta-vector, a new m-vector generated for a query via expansion performs about the same. These observations indicate that query expansion using retrieval feed-back is optimal when applied to the MeSH field. When this is done, no further improvement can be achieved by adding new terms to the free-text field of the query.

Comparison with Related Query-expansion Efforts

In comparison with the approach of Aronson et al.,[4,5] the improvements achieved using retrieval feedback are much higher than the 4% improvement in precision that they report.[4] Unfortunately, since Aronson et al. do not report their baseline scores, more informative comparisons cannot be made.

Comparison with SAPHIRE's performance requires calculation of an 11-AvgP score from the data Hersh et al. present in their Table 2.[1] This table lists SAPHIRE's average recall and precision scores at the cutoff points 0% to 95% in steps of 5%. Using the method of optimistic interpolation, we derive from this table an 11-AvgP score of 0.3549. This 0.3549 score, representing SAPHIRE's performance, is substantially lower than all of our baseline scores, suggesting that even without query expansion, retrieval using SMART is substantially more effective than retrieval using SAPHIRE. The comparison is a rough one, given that the recall and precision scores in their table appear to be micro averages computed on the total output for all queries. That is, their average precision at a recall point is the total number of relevant documents retrieved by all queries at that recall point divided by the total number retrieved over all queries. Instead, we computed macro averages. These methodologic variances suggest that the computed difference in performances is a somewhat tentative indicator of the relative merits of the approaches.

Comparison with Yang and Chute's work[7-9] is difficult since they used only a subset of the MEDLINE test collection that was used here. Differences in the distributions of relevant and nonrelevant documents might also influence such a comparison.

To conclude, our results indicate that query expansion based on retrieval feedback produces significant performance improvements for the MEDLINE database. Since this method does not require any prior relevance decisions, there is neither an extra cognitive load on the current user nor any reliance on prior users. Although the metathesaurus offers an extremely rich vocabulary, it appears that further innovation and research are needed to make profitable use of this global medical thesaurus for information retrieval. As pointed out by others, the mapping function from text to metathesaurus (or MeSH) is a crucial variable determining success. Given the current state of art, simple statistical and feedback methods, when combined with the SMART system's flexible index-term-weighting options, viably and effectively improve retrieval in MEDLINE.

Relevance Dependency of Retrieval Feedback

We examined query expansion via retrieval feedback to determine whether it was dependent on the presence of relevant documents in the feedback set, i.e., documents used for feedback information. If such dependence exists, then retrieval feedback closely resembles relevance feedback. Since relevance feedback is known to be highly effective, any significant improvement achieved with retrieval feedback would be no surprise. We examined this issue by determining performances on different subsets of queries where the subsets varied with respect to the proportions of their feedback documents that were actually relevant.

The best retrieval run within each expansion strategy specified in Table 6 was used for this analysis. For each selected run, the collection of 75 queries was partitioned into six non-overlapping subsets. The subsets defined queries with their feedback sets as 0%, 1–20%, 21–40%, . . . , and 81–100% relevant. For example, the best retrieval run under ES:ta'-m' is represented by the second row of Table 6. This run used a feedback set of five documents. Thus, the 75 queries are partitioned into those with 0, 1, 2, 3, 4, and 5 relevant documents in their feedback sets.

Results

Table 7 shows the results of expansion on query subsets for each retrieval run. It also presents the number of queries in each subset, and the baseline performance for each subset. The 0%-relevance subset has unique characteristics in that the percentage changes are remarkably high. Perhaps the extremely low baselines of these initial queries enable such conspicuous improvements. But despite these massive improvements, the final 11-AvgP scores remain quite low, suggesting that the quality of the original query places a ceiling on the performance of the final query.

The 1–20% subset is unique in that it is the first subset with somewhat reasonable baselines. The original queries yield on the average, a 11-AvgP score of 0.43 across the three retrieval-expansion combinations. It is also unique in that query-expansion strategies ES: ta'-m' and ES:ta' do not produce significant changes. This suggests that a feedback set that is 1–20% relevant is not sufficient for these two strategies to boost queries that start with somewhat adequate baseline performances. Beginning with the 21–40% subset, all strategies provide good improvements over baselines. The one exception is ES:ta' on the subset 81–100%, where performance declined slightly. In fact, the ES: ta' data indicate that for at least two subsets the added free-text terms may be too general, resulting in a poorer ranking of retrieved items. This result may explain why in our first experiment we found ES:ta' to be the weakest of the three alternative expansion strategies tested.

The comparatively smaller performance improvements achieved in the 81–100% subset may be because the baselines for the unexpanded queries are already quite high, leaving little room for improvement. Finally, if we focus our attention on the best expansion alternative, i.e., ES:ta-m', Table 7 suggests that for a new, original query one may expect anywhere from 9.3% to 119% improvement, depending upon its quality.

The 0% subset is the most pertinent one for testing the dependence of the query-expansion technique on relevant documents. This subset contains documents that are similar to the queries but not relevant to the queries. Since this subset yields the largest percentage improvements, we conclude that retrieval feedback does not depend upon the availability of relevant documents for feedback information. It appears that the vocabulary deficiency for these poorly constructed user queries is so large that similar but nonrelevant documents are capable of contributing effective terms.

Next, if we ignore the 0% subset, Table 7 reveals only a slight tendency for performance improvements to increase as the proportion of relevant documents increases. Percentages increase in the 1–60% range for two of the three retrieval runs. This tendency for improvements to increase appears to be dampened by the larger baselines of the higher-proportion subsets. This may explain why queries of highest quality produce smaller improvements than queries of medium quality.

Thus, we conclude that the query-expansion technique is effective even in the absence of relevant documents for feedback. Moreover, there is only a slight tendency for performance to improve with the availability of relevant documents.

Conclusions

Query-expansion strategies are needed to improve users' original queries to search the MEDLINE database. In this research, alternative query-expansion strategies within the retrieval-feedback framework were compared. The experiment indicated that the optimal strategy is one that adds only MeSH concepts to the original free-text queries. This increases the base-line 11-AvgP score averaged across all queries by 16.4% to 0.6018. Analysis of query subsets shows that, depending on the quality of the original free-text query, this expansion strategy offers improvements in the range of 9.3% to 119%. Additional expansion via free-text terms does not yield any further improvement. That is, expansion on the MeSH field appears

Table 7 ■

Performances with Query Subsets

	Percentage of Feedback Set That Is Relevant					
Expansion Strategy	0	1–20	21–40	41–60	61–80	81–100
Baseline	0.0721	0.4352	0.5159	0.5997	0.7418	0.8166
ES:ta'-m' (atc.atc)	0.1154	0.4468	0.5951	0.6851	0.8819	0.9149
Change	(+60.1%)	(+2.7%)	(+15.4%)	(+14.2%)	(+18.9%)	(+12%)
Number of queries	9	22	12	15	7	10
Baseline	0.0252	0.4246	0.5085	0.6316	0.6726	0.8642
ES:ta-m' (atn.atc)	0.0552	0.4737	0.5710	0.7648	0.8025	0.9441
Change	(+119%)	(+11.6%)	(+12.3%)	(+21.1%)	(+19.3%)	(+9.3%)
Number of queries	8	17	15	19	11	5
Baseline	0.0201	0.4199	0.5927	0.6413	0.6826	0.8460
ES:ta' (atn.atc)	0.0262	0.4189	0.6327	0.7597	0.7807	0.8384
Change	(+30.3%)	(−0.2%)	(+6.8%)	(+18.5%)	(+14.4%)	(−0.9%)
Number of queries	7	23	18	15	9	3

to overshadow any improvement that may be obtained by expanding the free-text field. Interestingly, the query expansion that totally ignored MeSH gave the lowest returns, and, in fact, reduced the best 11-AvgP score from 0.6018 to 0.5619 (i.e., by 6.63%), which could be a costly price to pay. Finally, expansion via retrieval feedback does not require the availability of relevant documents for feedback information.

Although our results are extremely encouraging, it should be noted that the test database used was small. This experimental feature limits the generalizability of our results. Having successfully designed expansion techniques using the small database, our next goal is to examine the scalability of these techniques to larger real-world databases. Thus, we are currently evaluating these techniques on the OHSUMED test collection,[2] which contains approximately 350,000 MEDLINE documents. However, the successful performance of related query-expansion techniques using the gigabyte-sized TREC databases[19] furthers our confidence in our techniques.

A second limitation of this study is that it utilized only Ide's method for feedback. Other feedback methods, such as Rocchio's,[17] might yield different results. Experimentation to investigate this dimension is planned for the future.

Although not a limitation, it should be noted that the majority of commercial retrieval systems for MEDLINE searching are Boolean retrieval systems. They do not retrieve documents by ranking as is done by SMART. Hence, the techniques investigated here are not directly applicable to such commercial retrieval systems, but appropriate modified versions may be easily developed and tested. Since users would appreciate ranked outputs over unranked ones, it is hoped that results offered here and by other explorations of retrieval by ranking will induce designers of the next generation of commercial retrieval systems to offer a choice of Boolean and ranking retrieval strategies.

Finally, different authors have recently debated, in various ways, the importance of MeSH for retrieval. The results obtained in this study underline the importance of MeSH for retrieval. Moreover, this research shows how to use the MeSH field successfully without requiring the user to be trained in the construction or the selection of MeSH concepts.

The author thanks the late Professor Salton, Amit Singhal, Chris Buckley, and Mandar Mitra for their comments on this paper and their generous hospitality. The author also thanks Professor Hersh for generously providing the MEDLINE test database.

References ■

1. Hersh W, Hickam D, Haynes R, McKibbon K. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. J Am Med Informatics Assoc. 1994;1:51–60.
2. Hersh W, Buckley C, Leone T, Hickam D. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: Croft B, van Rijsbergen C, eds. SIGIR 1994: Proceedings of the 1994 International Conference on Research and Development in Information Retrieval. New York: ACM, 1994:192–200.
3. Srinivasan P. Exploring query expansion strategies for MEDLINE. Technical Report, Department of Computer Science, Cornell University, Ithaca, NY 14853, 1995.
4. Aronson A, Rindflesch T, Browne A. Exploiting a large thesaurus for information retrieval. In: Brentano P, Seitz F, eds. RIAO 1994: Proceedings of the 1994 RIAO Conference. New York: ACM, 1994:197–216.
5. Rindflesch T, Aronson A. Ambiguity resolution while mapping free text to the UMLS metathesaurus. In: SCAMC 1994: Proceedings of the 1994 Symposium on Computer Applications in Medical Care. J Am Med Informatics Assoc. 1994; 1(suppl):240–4.
6. Voorhees E, Hou Y-W. Vector expansion in a large collection. In: Harman D, ed. TREC-1: Proceedings of the First Text REtrieval Conference. Washington, DC: Government Printing Office, 1992:343–51.
7. Yang Y. Expert Network: effective and efficient learning from human decisions in text categorization and retrieval. In: Croft W, van Rijsbergen C, eds. SIGIR 1994: Proceedings of the 1994 International Conference on Research and Development in Information Retrieval. New York: ACM, 13–22.
8. Yang Y, Chute C. Words or concepts: the features of indexing units and their optimal use in information retrieval. In: Safran C, ed. ŚCAMC 1993: Proceedings of the 1993 Symposium on Computer Applications for Medical Care. New York: McGraw-Hill, 685–9.
9. Yang Y, Chute C. An application of least squares fit mapping to text. In: Korfhage R, Rasmussen E, Willett P, eds. SIGIR 1993: Proceedings of the 1993 International Conference on Research and Development in Information Retrieval. New York: ACM, 281–90.
10. Crouch C. An approach to the automatic construction of global thesauri. Information Processing and Management. 1990;26:629–40.
11. Crouch C, Yang B. Experiments in automatic statistical thesaurus construction. In: Belkin N, Ingwersen P, Pejtersen A, eds. SIGIR 1992: Proceedings of the 1992 International Conference on Research and Development in Information Retrieval. New York: ACM, 1992:77–88.
12. Jing Y, Croft W. An association thesaurus for information retrieval. In: Brentano P, Seitz F, eds. RIAO 1994: Proceedings of the 1994 RIAO Conference. New York: ACM, 1994: 146–60.
13. Schutze H, Pedersen J. A co-occurrence-based thesaurus and two applications to information retrieval. In: Brentano P, Seitz F, eds. RIAO 1994: Proceedings of the 1994 RIAO Conference. New York: ACM, 1994: 266–74.
14. Sparck Jones K, Needham R. Automatic term classification and retrieval. Information Processing and Management. 1968;4:91–100.
15. Buckley C, Salton G. Optimization of relevance feedback weights. To appear in SIGIR 1995: Proceedings of the 1995 International Conference on Research and Development in Information Retrieval (SIGIR 1995). New York: ACM, 1995.

16. Ide E. New experiments in relevance feedback. In: Salton G, ed. The SMART Retrieval System—Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice-Hall, 1971:337-54.
17. Rocchio J. Relevance feedback in information retrieval. In: Salton G, ed. The SMART Retrieval System—Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice-Hall, 1971:68-73.
18. Salton G, Buckley C. Improving retrieval performance by relevance feedback. J Am Soc Info Sci. 1990;41:288-97.
19. Harman D. Overview of the third Text REtrieval Conference (TREC-3). In: Harman D, ed. TREC-3: Proceedings of the Third Text REtrieval Conference. Washington, DC: Government Printing Office, 1994:1-19.
20. National Library of Medicine. Unified Medical Language System (UMLS) Knowledge Sources, 5th experimental edition. Bethesda, MD: NLM, 1994.
21. Elkin P, Cimino J, Lowe H, et al. Mapping to MeSH. In: Greenes RA, ed. SCAMC 1988: Proceedings of the 1988 Symposium on Computer Applications for Medical Care. Washington DC, 1988:185-190.
22. Buckley C. Implementation of the SMART information retrieval system. Technical Report TR85-686, Department of Computer Science, Cornell University, Ithaca, NY 14853, May 1985.
23. Salton G, ed. The SMART Retrieval System—Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice-Hall, 1971.

APPENDIX A

Example Relevant Document

Document number 125, ranked 33 by original query and 9 by expanded query.

Title: A 10-year experience with topical mechlorethamine for mycosis fungoides: comparison with patients treated by total-skin electron-beam radiation therapy.

Authors: Vonderheid EC; Van Scott EJ; Wallner PE; Johnson WC

Abstract: A group of 243 patients with mycosis fungoides (MF) received treatment with topical applications of dilute aqueous solutions of mechlorethamine and/or systemic chemotherapy over the past 10 years. The likelihood of a complete and relapse-free remission and survival was found to correlate inversely to the magnitude of disease as denoted by a simple staging system. Although disease-free intervals of greater than 3 years have occurred thus far in 32 (13permanency of these remissions and the curability of disease remain uncertain because of the variability of disease progression characteristic of MF. Comparison of treatment results with those published on a large group of patients treated with total-skin electron-beam radiation therapy indicates that the chemotherapeutic approach to the treatment of MF is equally effective in promoting survival.

MeSH entries:

Administration, Topical
Adolescence
Adult
Aged
Child
Comparative Study
Female
Human
Male
Mechlorethamine/*ADMINISTRATION & DOSAGE
Middle Age
Mycosis Fungoides/*THERAPY
Radiotherapy, High Energy
Remission, Spontaneous
Sezary Syndrome/THERAPY
Skin Neoplasms/*THERAPY
Support, U.S. Gov't, P.H.S.
Time Factors

APPENDIX B

Example Non-relevant Document

Document number 240, ranked 29 by original query and 45 by expanded query.

Title: E-rosette inhibitory factor in sera from patients with mycosis fungoides.

Abstract: Peripheral blood lymphocytes from some of patients with mycosis fungoides disease showed decreased ability to form rosettes with sheep erythrocytes. This decreased percentage of E-rosette forming cells could be normalized when those cells were incubated in culture for 20 hr. Since these data led us to considering a possible inhibitory factor present in patients' sera, we tested their ability to inhibit E-rosetting by T lymphocytes from normal donors, and found that sera from mycosis fungoides patients with low levels of E-rosetting blood lymphocytes showed greater inhibitory effect on E-rosette formation by normal T cells when compared to those either from normal donors or from mycosis patients who had almost normal levels of E-rosetting blood lymphocyte number. The E-rosette inhibitory factor was sensitive to 2-mercaptoethanol treatment and was copurified with serum IgM by ammonium sulfate precipitation and by sequential gel filtrations, suggesting that it might be an anti-T lymphocyte antibody naturally occurring during the disease process.

MeSH entries:

Adult
Aged
Animal
Antilymphocyte Serum/ISOLATION & PURIF
Blood Proteins/*ISOLATION & PURIF
Female
Human
Male
Mercaptoethanol/PHARMACOLOGY
Middle Age
Mycosis Fungoides/*IMMUNOLOGY
Rosette Formation
Sheep/IMMUNOLOGY
T-Lymphocytes/*IMMUNOLOGY